АНАЛІЗ ПРОБЛЕМИ АВТОМАТИЧНОГО ВИЯВЛЕННЯ ПЕРЕФРАЗУВАНЬ В ОБЛАСТІ ЗАВДАНЬ NATURAL LANGUAGE PROCESSING

Петрасова С.В.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Автоматичне виявлення перефразувань є одним з актуальних завдань Natural Language Processing (NLP), що зумовлене труднощами врахування як синтаксичних, так і семантичних зв’язків між словами.

Інтерес до виявлення перефразувань, зокрема семантично близьких колокацій, сприяв формуванню різних підходів до їх видобування.

Семантико-синтаксичний підхід передбачає аналіз колокацій в аспекті теорії стійких сполучень і граматики конструкцій. Колокації розглядаються як комплексні семантико-синтаксичні одиниці, які характеризуються семантичною, синтаксичною і дистрибутивною регулярністю.

Перевагою корпусно-орієнтованого підходу є використання статистико-лінгвістичного апарату корпусу текстів для виявлення релевантних граматично правильних і семантично значущих колокацій. Розвиток корпусно-орієнтованого підходу забезпечується використанням великих обсягів корпусів текстів, вбудованих програм лематизації, морфологічних та синтаксичних фільтрів і латентного семантичного індексування [1].

Автоматичне видобування перефразувань зазвичай виконується як процес, що складається з декількох етапів. На першому етапі виявлення колокацій застосовуються статистичні міри (міри асоціації), що обчислюють силу зв’язку між елементами в складі колокації та враховують як частоту спільної зустрічальності, так і частоту кожного окремого елемента в корпусі текстів. Наступний етап ідентифікації семантично близьких колокацій потребує використання спеціалізованих лінгвістичних ресурсів, наприклад, словників синонімів або тезаурусів.

В результаті проведеного аналізу підходів та методів автоматичного видобування перефразувань (семантично близьких колокацій) пропонується здійснити комбінацію лінгвістичних і статистичних критеріїв. Спочатку відбираються словосполучення, що відповідають певним лінгвістичним критеріям, а потім отриманий список скорочується за допомогою статистичних критеріїв. Саме комбінування різних методів складає основу успішного розв’язку проблем, пов’язаних з NLP.

Література:

1 Бобкова Т.В. Теоретико-методологічні підходи до вивчення колокацій у сучасному мовознавстві. Вісник КНЛУ. Серія: Філологія. Т. 17. № 2. 2014. С. 14–22.